Computational Statistics and Data Analysis 103 (2016) 413–425

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Semiparametric mixture: Continuous scale mixture approach

Sijia Xiang [a], Weixin Yao [b], Byungtae Seo [c,*]

[a] School of Mathematics and Statistics, Zhejiang University of Finance & Economics, Hangzhou, Zhejiang 310018, PR China

[b] Department of Statistics, University of California, Riverside, CA 92887, USA

[c] Department of Statistics, Sungkyunkwan University, Seoul, Republic of Korea

ARTICLE INFO

Article history:
Received 30 October 2015
Received in revised form 2 June 2016
Accepted 2 June 2016
Available online 11 June 2016

Keywords:
Mixture models
Semiparametric EM algorithm
Semiparametric mixture models
Continuous normal scale mixture

ABSTRACT

In this article, we propose a new estimation procedure for a class of semiparametric mixture models that is a mixture of unknown location-shifted symmetric distributions. The proposed method assumes that the nonparametric symmetric distribution falls in a rich class of continuous normal scale mixture distributions. With this new modeling approach, we can suitably avoid the misspecification problem in traditional parametric mixture models. In addition, unlike some existing semiparametric methods, the proposed method does not require any modification or smoothing of the likelihood as it can directly estimate parametric and nonparametric components simultaneously in the model. Furthermore, the proposed parameter estimates are robust against outliers. The estimation algorithms are introduced and numerical studies are conducted to examine the finite sample performance of the proposed procedure and to compare it with other existing methods.

1. Introduction

Finite mixture models have a wide application including cluster and latent class analysis, discriminant analysis, image analysis, and survival analysis. They provide extremely flexible descriptive models for distributions in data analysis and inference. For general introduction of mixture models, see Lindsay (1995), Böhning (1999), McLachlan and Peel (2000) and Frühwirth-Schnatter (2006).

A general form of a finite mixture density can be expressed as

$$p(x;\boldsymbol{\theta}) = \pi_1 f(x;\lambda_1) + \pi_2 f(x;\lambda_2) + \cdots + \pi_m f(x;\lambda_m),$$

where $\boldsymbol{\theta} = (\lambda_1, \ldots, \lambda_m, \pi_1, \ldots, \pi_m)$, $\sum_{j=1}^{m} \pi_j = 1$, $\pi_j > 0$ for $j = 1, \ldots, m$, and $f(x;\lambda_j)$ is the density function for the jth component. The traditional parametric mixture model assumes that the density f belongs to a certain parametric family, such as family of normal distributions or t-distributions. The maximum likelihood estimator (MLE) of the unknown parameter $\boldsymbol{\theta}$ can be then obtained using the Expectation–Maximization (EM) algorithm.

In practice, however, a practitioner might not have prior information about which parametric family one should use for f. The estimate of $\boldsymbol{\theta}$ might be sensitive to the parametric form of f, and in addition, a distributional misspecification of f could lead to wrong or inefficient statistical inference. For this, one can consider a semiparametric model that leaves f completely unspecified. Yet, this causes an identifiability problem as the model is too flexible and not parsimonious enough. For this

* Corresponding author.
E-mail address: seobt@skku.edu (B. Seo).

http://dx.doi.org/10.1016/j.csda.2016.06.001
0167-9473/

identifiability issue in semiparametric mixture models, Bordes et al. (2006) and Hunter et al. (2007) considered the following location-shifted semiparametric model with symmetric nonparametric component densities:

$$p(x;\boldsymbol{\theta},f)=\sum_{j=1}^{m}\pi_j f(x-\mu_j), \tag{1.1}$$

where $\boldsymbol{\theta}=(\pi_1,\mu_1,\ldots,\pi_m,\mu_m)$ and f is an unknown but symmetric density about zero. Bordes et al. (2006) proved the identifiability of model (1.1) for $m=2$. Hunter et al. (2007) further established the identifiability of model (1.1) for both $m=2$ and $m=3$.

Bordes et al. (2007) and Benaglia et al. (2009) proposed a semiparametric EM type algorithm to estimate parameters in (1.1) using a kernel-based estimator for the symmetric nonparametric density f. They demonstrated, through numerical study, its superiority over the methods provided by Hunter et al. (2007) and Bordes et al. (2006). However, the bandwidth selection is not an easy task and sensitive to the model efficiency. In this case, many commonly used methods for bandwidth selection may not be relevant because each component density may have a different impact on the choice of bandwidth and the ideal bandwidth selection depends on whether components are well-separated or not.

In this article, we propose a new method to estimate the model parameters in (1.1) by modeling f as nonparametric scale mixtures. The proposed method is free from bandwidth selection and thus is more reliable and robust to model misspecification. Unlike Bordes et al. (2007) or Benaglia et al. (2009), the new technique relies only on the likelihood function without any modification or smoothing. In addition, it can give a direct legitimate nonparametric estimator of f. Furthermore, the proposed parameter estimates are robust against outliers.

The remainder of this paper is organized as follows. In Section 2, we introduce the new estimation method for the semiparametric mixture model (1.1) and an effective algorithm is introduced to find the proposed estimator in Section 3. In Section 4, we present both a Monte Carlo study and a real data example to compare the proposed new estimator with some other existing methods. Finally, some discussions are given in Section 5.

2. Semiparametric mixtures under continuous scale mixture

For the nonparametric symmetric density f in (1.1), we propose to model f as a continuous normal scale mixture. That is, we assume that f is a member of

$$\mathcal{F}=\left\{f(x)|\int\frac{1}{\sigma}\phi(x/\sigma)dQ(\sigma)\right\}, \tag{2.1}$$

where $\phi(x)$ is the standard normal density, and Q is an unspecified probability measure on $\mathbb{R}^+$. Although we restrict the nonparametric symmetric density f to $\mathcal{F}$, $\mathcal{F}$ is rich enough to contain almost all symmetric unimodal continuous probability densities such as normal, Laplace, t, stable, and so on. Efron and Olshen (1978) and Basu (1996) discussed on how many distributions are contained in $\mathcal{F}$. Kelker (1971) and Andrews and Mallows (1974) also studied necessary and sufficient conditions for a probability density to be a member of $\mathcal{F}$. Recently, Seo and Lee (2015) utilized this class of normal scale mixture densities to efficiently estimate the distribution of innovations as well as parameters in semiparametric generalized autoregressive conditional heteroskedasticity models. Böhning and Ruangroj (2002) discussed the difference between the normal with a free variance parameter and component mixture of normals with the same mean for $m=2$. Böhning and Ruangroj (2002) proved in Theorem 2.3, the difference is an increasing function of the contaminated component variance when other parameters are fixed. However their results are limited to two-component normal scale mixtures and thus cannot be applied to the continuous normal scale mixtures. Since the t-distribution is a special case of the continuous normal scale mixtures, the difference can be very big between normal distribution with a free variance parameter and the continuous normal scale mixtures.

Under this model class, (1.1) can be expressed as

$$\begin{aligned}p(x;\boldsymbol{\theta},Q)&=\sum_{j=1}^{m}\pi_j\left\{\int\frac{1}{\sigma}\phi\left(\frac{x-\mu_j}{\sigma}\right)dQ(\sigma)\right\}\\&=\int\sum_{j=1}^{m}\frac{\pi_j}{\sigma}\phi\left(\frac{x-\mu_j}{\sigma}\right)dQ(\sigma).\end{aligned} \tag{2.2}$$

The identifiability of (2.2) can be shown by combining the identifiability of (1.1) and $\mathcal{F}$ as described in Proposition 2.1. Chee and Wang (2013) used a similar argument for the identifiability of their semiparametric location mixtures.

Proposition 2.1. *The semiparametric model $p(x;\boldsymbol{\theta},Q)$ in (2.2) is identifiable when $m\le 3$, i.e., if $p(x;\boldsymbol{\theta},Q)=p(x;\boldsymbol{\theta}^*,Q^*)$, then $Q=Q^*$ and $\boldsymbol{\theta}=\boldsymbol{\theta}^*$ up to a permutation of component labels.*